

PySpark Runtime Architecture

In this document, we'll explore the
PySpark Runtime Architecture.

[LINKEDIN.COM/IN/ARUDSEKABERNE/](https://www.linkedin.com/in/arudsekaberne/)



Submit our application into Cluster

Using Spark Submit Command



Spark Submit Command

```
spark-submit \  
  --master yarn \  
  --deploy-mode cluster \  
  --executor-cores 4 \  
  --num-executors 2 \  
  --executor-memory 16G \  
  --driver-memory 16G \  
  /path/to/PySpark.py
```

YARN RM

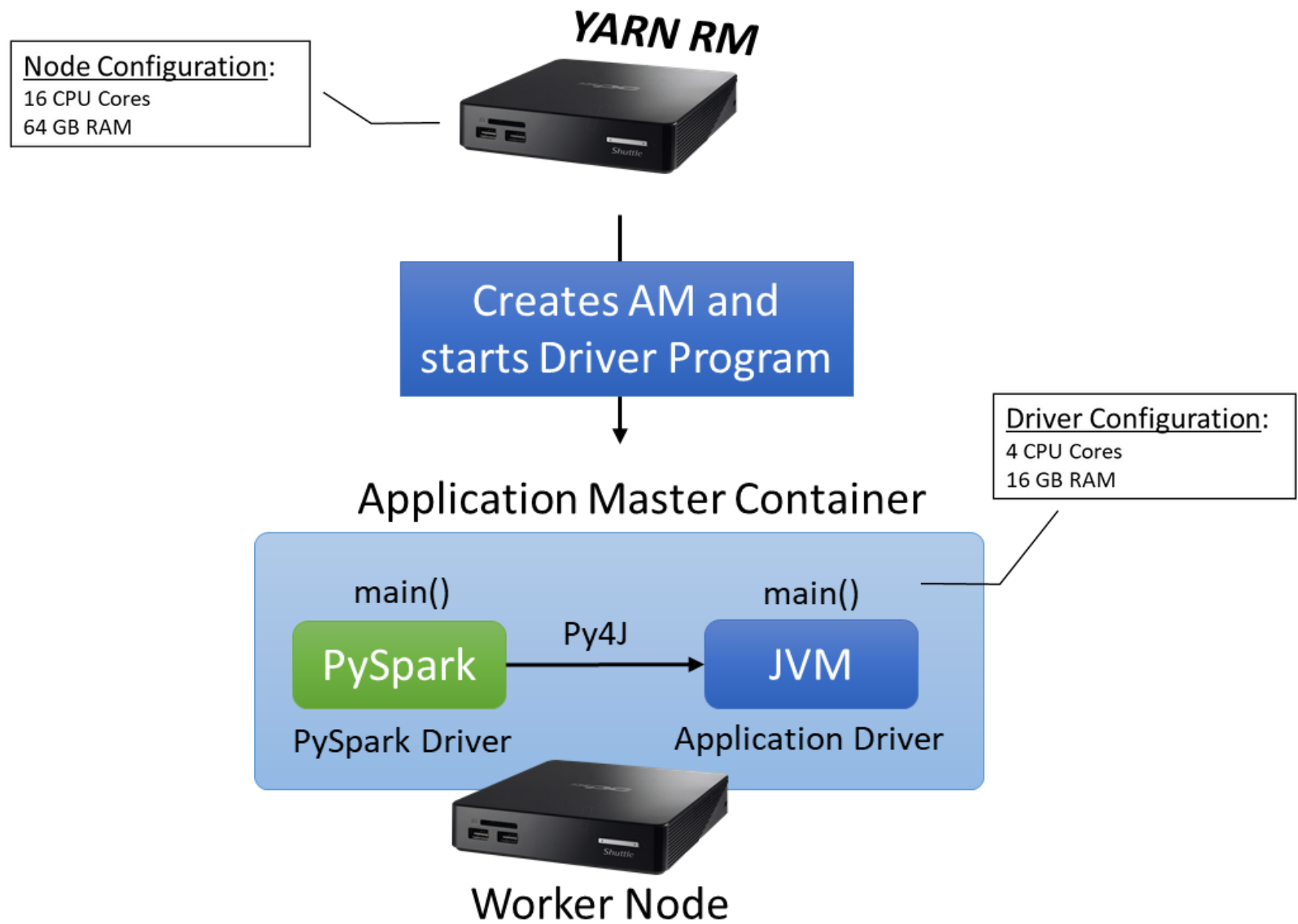


4 node cluster

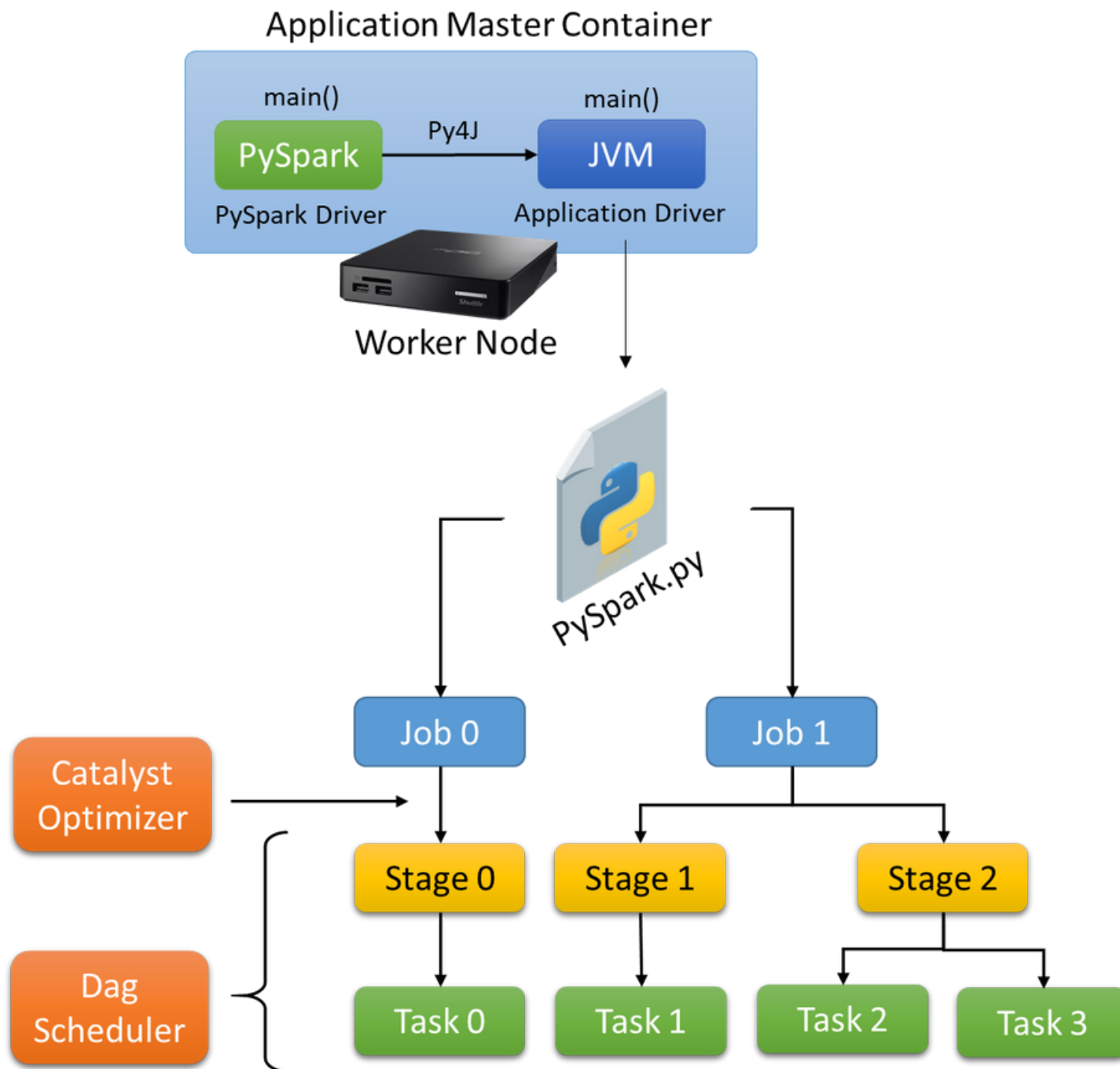




What happens when we submit our application into cluster?

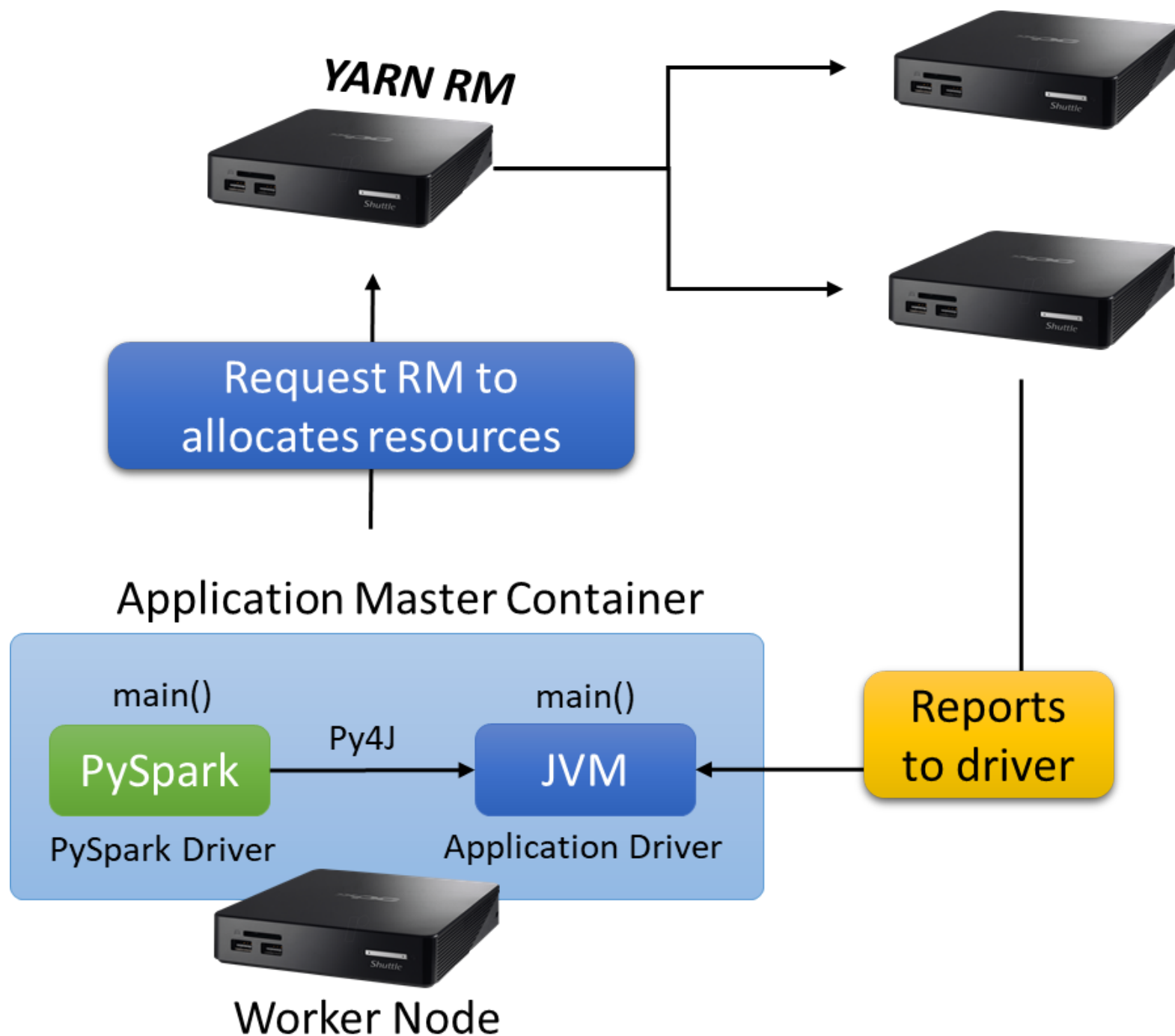


Driver compiles code into multiple Jobs based on Action and
Dag Scheduler create stages based on Wide Transformation
respectively



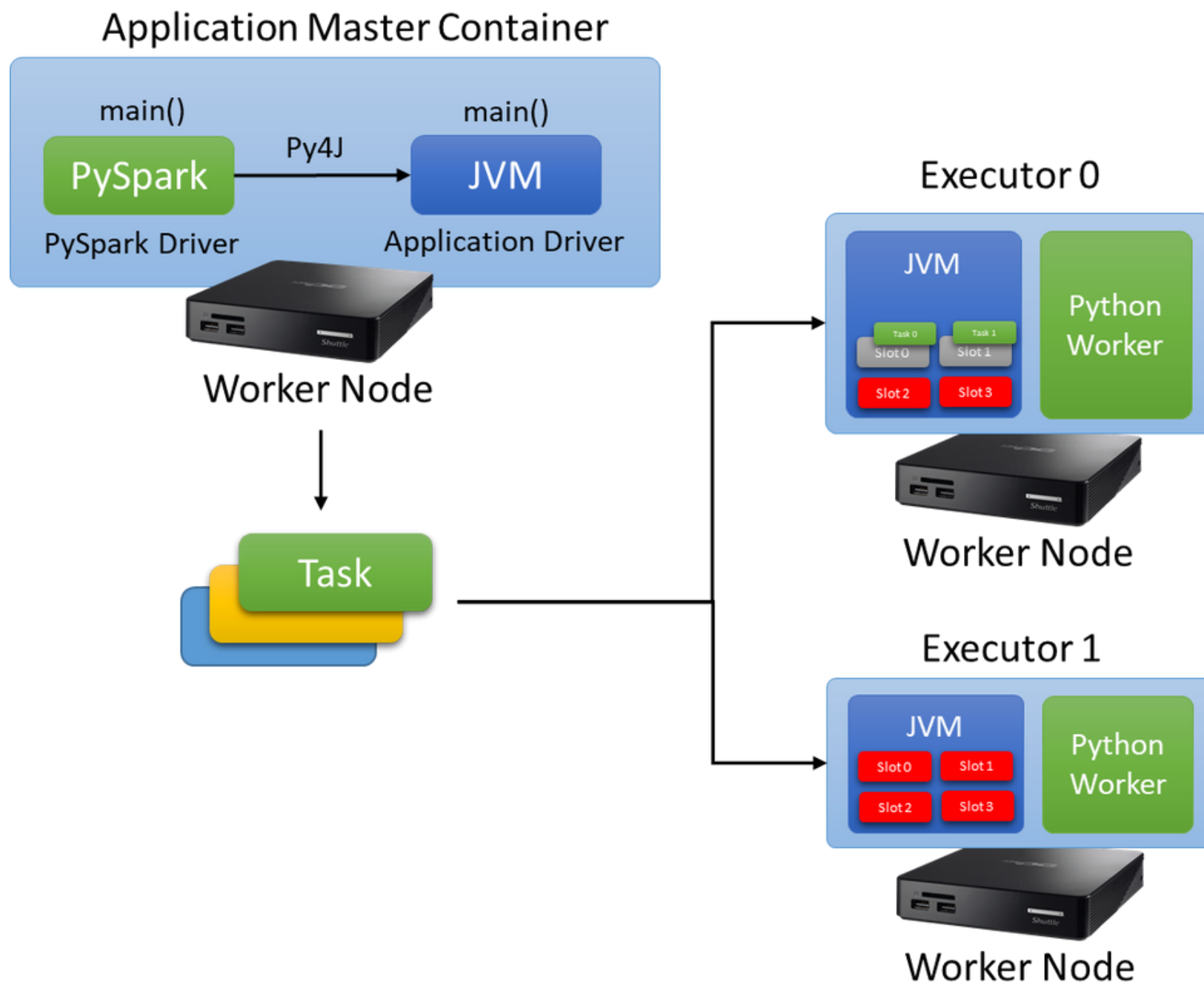
Driver Request Resource Manager

to create requirement resource for job execution



Driver creates and launches the Executors

Executors perform the actual processing of the data by executing tasks assigned to them by the driver



Have you enjoyed this overview of PySpark's
Runtime Architecture?



Arud Seka Berne S

Follow me

Follow me for more in-depth technical content on
big data and related topics.